PIRSF Names: Standards

Origin:

The ***main source*** of PIRSF names will come from the literature, to identify the most common among various alternatives. Alternative names for the same activity used with near-equal frequency can be identified with the alternative in parentheses in the name field, but otherwise such can be provided within the description or an "alternative names" field.

e.g. NADH:ubiquinone oxidoreductase, subunit 2 (chain N)

SF names should take into account any published standardizations. Often these can be found by a literature search for "unified nomenclature". If no accepted unification exists, and several alternatives are of equal frequency in the literature, use the one with the easiest extensibility or standardization (among subunits, for example).

Names provided by the Enzyme Commission should be used whenever possible, unless there is a conflict with more prevalent forms in literature.

e.g. EC uses leucine-tRNA ligase, while the predominant form used in literature (and by Swiss-Prot) is leucyl-tRNA synthetase.

Please see **Curation Notes** for more detailed information on where to look for names.

Abbreviations:

In cases where the function is commonly referred to by an abbreviation, indicate it in the name.

e.g. Transport protein particle (TRAPP) complex subunit

Abbreviations of chemicals, such as Ca or Mg, are allowed. However, certain chemicals may be spelled out under some conditions. A list of standards for various cases will be created.

Common Phrasing:

"uncharacterized conserved protein": Used for PIRSFs with no revealing information.

"involved in": In some cases, the exact function of a protein is not yet known, but its involvement in a pathway or complex may be known or inferred. In such cases, it is suitable to use the phrase "…involved in…" to indicate this information.

Relatedness:

If no information is known other than the relatedness of one protein to another (hopefully of known function), this can be indicated by saying "…related to…". Avoid hyphenating if possible, since this can get very unwieldy.

e.g. uncharacterized endonuclease related to Holliday junction resolvase (OK)
Holliday junction resolvase-related endonuclease (no)

In cases where the PIRSF is composed of likely-inactive proteins similar to enzymes, use the word homolog.

e.g. inactive protease homolog

Also see notes under **Qualifiers** below.

Location:

Often nothing is known about a protein other than its known or predicted location in the cell. If the location is membrane, this can be indicated in the name. Cytoplasmic or nuclear location probably is not worthwhile indicating.

For membrane proteins, when nothing else is known about the PIRSF, the phrase "membrane protein" should be used. When membrane location has been experimentally determined, the TENTATIVE tag should be used; otherwise, use the PREDICTED tag. Also, the word "membrane" in the name can be modified to "transmembrane" or "integral membrane" as appropriate. In all cases of names consisting solely of location, it is likely most appropriate to use a short name as is done for uncharacterized proteins: "UCP <PIRSF number>". For example, for PIRSF109898, the short name would be UCP109898.

Uniqueness:

Non-unique PIRSF names (aka short descriptions) are permitted.

Case:

Names are lowercase, except for commonly-capitalized abbreviations.

e.g. DNA gyrase
e.g. leucyl-tRNA synthetase

Functionally Non-Homogeneous PIRSFs:

The name should be appropriate for the whole PIRSF, so that one can read the name as "each member of this PIRSF is <insert name here>". In some cases, this will require that the name given be of a general type, when several members have been characterized to have the general activity but with different substrates. Further explanation will be put in the description. Also, a box indicating that the PIRSF is composed of proteins of various functions will be checked. This will alert users that there may be a more appropriate (more specific) name upon further checking.

e.g. sugar transporter
(perhaps contains glucose, mannose, fructose transporters)

NOTE: This situation differs from the case of PIRSFs composed of proteins that have multiple functions within a single polypeptide. In such cases, the box is NOT checked.

Functionally Homogeneous Multifunctional PIRSFs:

Certain PIRSFs are comprised of proteins that are each multifunctional. The name should reflect this situation (usually, such cases are named accordingly in the literature

anyway). PIRSFs composed of bifunctional proteins are a special class that gets a "slash" between each
function. No spaces should surround the slash.

Non-propagatable names:

In (hopefully) rare cases, it is not possible or advisable to give an PIRSF a name that is general enough so that it is suitable for every protein. That is, the name fails the "each member of this PIRSF is <PIRSF name>" test. In such cases, a tag (Non-propagatable) can be checked to ensure that the PIRSF name is not put onto any individual protein. Checking this tag would be tantamount to indicating that "each member of this PIRSF is a member of the <PIRSF name> family". IT IS EXPECTED THAT EVERY EFFORT WILL BE MADE TO DEVISE A GENERAL NAME BEFORE USING THIS TAG, and that ONLY A VERY SMALL MINORITY OF PIRSFs WILL BE THUS TAGGED. Note also that the non-propagate tag is NOT meant to indicate the goodness of the name--it is only meant to indicate the complete unsuitability of the name as applied to an individual protein.

Checkboxes for validated, tentative, and predicted:

"Validated" means at least one protein in the PIRSF has been experimentally characterized to have the activity indicated by the name. If there are multiple activities, and one is validated and the other is predicted, use "validated" and put an explanation in the description.

"Tentative" means that there is experimental evidence for the name given for the PIRSF, but the evidence is not really conclusive.

"Predicted" will be used for *any* type of prediction of *any* confidence level. Evidence for the prediction will go in the description.

NOTE: These three boxes are mutually exclusive. However, they can be combined with the "non-homogeneous function" box described above.

Qualifiers:

If necessary, qualifiers to names can be used to distinguish one PIRSF from another. However, PIRSF names will primarily be used to name individual proteins, so keep this in mind when naming. The preferred type of qualifier would indicate function, but the following may also be used:

Gene/protein names: come after a comma. Use the protein form (for *E. coli*, uppercase first letter; for yeast, same form as coli but with "p" added to the end).

e.g. response regulator, AlgR type (correct, protein version)
response regulator, algR type (wrong, gene version)
origin recognition complex, subunit 1 (Orc1p) (protein version)
origin recognition complex, subunit 1 (ORC1) (in yeast, this is a gene)

Domain qualifiers: come after "with" (see below under **Domains/motifs** for more info). If an article (a/an/the) is required, avoid using "the."

e.g. protease IV with duplicated peptidase family U7 domain

Name Specificity:

The name should be as specific as possible while still upholding the requirement that the name be suitable for every protein.

e.g. response regulator, AlgR type
signal transduction protein, AlgR type (less specific)
response regulator AlgR (too specific—other members might have a different gene name)

Multisubunit complexes:

Named according to the complex, followed by a comma and the specific subunit name.

e.g. ribonucleotide reductase, alpha subunit

(This is to distinguish between a subunit of a complex and an unknown or unnamed component of certain classed enzyme).

e.g. ribonucleotide reductase alpha subunit (ambiguous)

Exception: when using the word “component” do not precede with a comma.

Search for papers describing standardizations/unified nomenclature.

Subunits:

The word “subunit” is preferred over “chain” or “component”. However, the predominant usage in literature always takes precedence.

Precede the word "subunit" with:

Greek letters	(ribonucleotide reductase, alpha subunit)
Size indicators	(3-isopropylmalate dehydratase, large subunit)
	(NADH:ubiquinone oxidoreductase, 20 kD subunit)
Protein IDs	(multisubunit Na+/H+ antiporter, MnhC subunit)
Type of subunit	(Serine/threonine protein phosphatase 2A, regulatory subunit)

Follow the word "subunit" with:

Letters	(F0F1-type ATP synthase, subunit a)
Numbers/Roman numerals	(anaphase-promoting complex (APC), subunit 10)

Domains/motifs:

In cases where nothing is known about the protein other than a domain or motif that doesn't offer enough clues as to function (most often will occur with uncharacterized proteins), add the phrase "with ..." after the given name. Avoid hyphenating as in “<some domain>-containing”.

e.g. uncharacterized conserved protein with RING Zn-finger (OK)
PAS domain-containing protein (please, no)

However, if the domain does give enough information to make a prediction, use the prediction as the name and cite the domain evidence in the description or comments.
e.g. nucleic acid-binding protein
{in description or comments, indicate that it has a PIN domain}

If there are several classes of a given motif, indicate it
e.g. RING Zn-finger
CCCH-type Zn-finger

It is preferable to use the full name of a domain as given in the literature rather than the abbreviated versions used in Pfam or SMART.

Hyphenation:

In enzyme names, use a single dash (-) between greek letters and numbers.
e.g. delta-9 acyl-ACP desaturase

Transfer enzymes often indicate the source and destination cofactors. One method of indicating such is with a colon. However, another method uses a double dash (--).
e.g. formylmethanofuran--tetrahydromethanopterin formyltransferase
(this enzyme transfers a formyl group from formylmethanofuran to tetrahydromethanopterin)

When indicating “type”, when used as a mechanism to differentiate between PIRSFs, do not use a hyphen.
e.g. 6-phosphofructokinase, eukaryotic type (OK)
6-phosphofructokinase, eukaryotic-type (incorrect)

(NOTE: bear in mind that some hyphens will occur before the word “type”
e.g. V-type ATPase)

Hyphens should typically be used before the word “dependent”
e.g. pyrophosphate-dependent phosphofructokinase

Special cases:

Sigma subunits should have type follow "sigma" with no space so the name can be automatically formatted to the correct superscripted form
e.g. RNA polymerase, sigma70 subunit

Charged tRNAs are indicated by "tRNA" followed by the three-letter amino acid code, with the first letter capitalized
e.g. Asp-tRNAAsn/Glu-tRNAGln amidotransferase, subunit B

cytochrome c:
- lowercase “c”
- no dash, unless it is a modifier for an enzyme
e.g. cytochrome-c reductase

“stand-alone” is indeed hyphenated

Avoid:

Specific sizes (in kDa)--these usually don't apply to a whole PIRSF--unless the size has been applied to many proteins that are not really that size. If no alternative exists, use the form “XX kDa” (note the space and the capitalization).

Indicating species or indicating "conserved in ..." (the former is too specific and the latter may be proven too specific later on).

EC numbers do not belong as part of a PIRSF name.

The word “precursor” unless the size is much larger than the mature peptide.

The use of the articles “a”, “an”, and “the”.

?????:

In general, how important is it to use names that give distinction to PIRSF names, as by indicating “,XXX type” or “, XXX family”?

How to refer to names when family or superfamily (as defined in literature, for example, a domain superfamily) affiliation is desired, given that external usage of these terms may differ from ours? Note that affiliation serves only to distinguish one PIRSF from another, and may not be critical.

When an additional domain distinguishes one PIRSF from another, do we want to indicate the distinction? We use “with” when the domain is the only notable feature, but do we want to use the same word, or something else. For example:

“sensor histidine kinase”
“sensor histidine kinase with an ABC domain”

When indicating relationship, which is preferred:

ATP-dependent protease related to Clp
ATP-dependent protease of the Clp type
ATP-dependent protease of the Clp family
ATP-dependent protease (Clp family)
ATP-dependent protease (Clp homolog)

PIRSF Short Names: Standards

"Short names" must be provided to InterPro. There is a 16 character limit to these, which includes all alphanumeric characters (upper and lower case), dashes, and underscores. Commas and spaces are not allowed. A small list of standards is available.

Short names for "uncharacterized conserved protein" or "membrane protein" (and relatives) consist of the prefix UCP followed (without spaces) by the PIRSF number. Therefore, the short name for PIRSF109898 (assuming it is uncharacterized or membrane, etc.) would be UCP109898.